Artificial intelligence method developed for classifying raw sugarcane in the presence of the solid impurity

Lucas Janoni dos Santos[1,2] , Érica Regina Filletti[1,2] , Fabiola Manhas Verbi Pereira[1,2,3+]

1. São Paulo State University, Institute of Chemistry, Araraquara, Brazil.
2. Bioenergy Research Institute, Group of Alternative Analytical Approaches, Araraquara, Brazil.
3. National Institute of Alternative Technologies for Detection Toxicological Assessment and Removal of Micropollutants and Radioactive Substances, Araraquara, Brazil.

+Corresponding author: Fabiola Manhas Verbi Pereira, Phone: +55 (16) 33019609, Email address: fabiola.verbi@unesp.br

ABSTRACT: An investigation dedicated to evaluating a big issue in biorefineries, solid impurity in raw sugarcane, is presented. This relevant industrial sector requests a high-frequency, low-cost, and noninvasive method. Then, the developed method uses the averaged color values of ten color-scale descriptors: R (red), G (green), B (blue), their relative colors (r, g, and b), H (hue), S (saturation), V (value) and L (luminosity) from digital images acquired from 146 solid mixtures among sugarcane stalks and solid impurity — vegetal parts (green and dry leaves) and soil. The solid mixture of samples was prepared considering desirable and undesirable scenarios for the solid impurity amounts. The outstanding result was revealed by an artificial neural network (ANN), achieving 100% of accurate classifications for two ranges of raw sugarcane in the samples: from 90 to 100 wt% and from 41 to 87 wt%. Low-computational cost and a simple setup for image acquisition method could screen solid impurity in sugarcane shipments as a promising application.

ANN MODEL FOR BIOENERGY
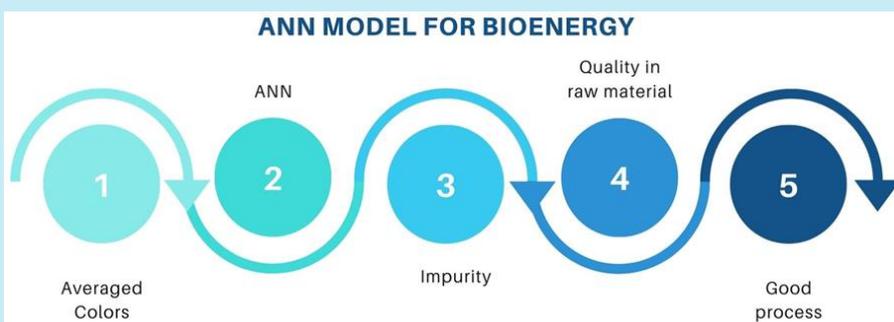
1. Introduction

Image and color information has played an important role in analytical chemistry and can help solve many issues, mainly because of its versatility and availability of many low-cost devices for *in loco* or laboratory analysis (Capitán-Vallvey et al., 2007; Diniz, 2020; Pereira and Bueno, 2007; Pereira et al., 2011).

Our research group has developed analytical methods to evaluate raw sugarcane to help the mills or biorefineries manufacturing process of this material routinely monitored as a consignment for payment purposes. The quality of raw sugarcane influences the manufacturing process, directly compromising two essential commodities — sugar and ethanol (Andrade et al., 2018; Guedes and Pereira, 2018; 2019; Guedes et al., 2020; Romera et al., 2016).

Solid impurity in raw sugarcane is defined as the plant presence (tops, green, brown, and dry leaves) and the soil (Eggleston et al., 2010). This issue is impacted by the type of harvesting process, as harvesting green or burnt cane. In specific, harvesting green increases the quantity of solid impurity in raw sugarcane, as reported in technical notes and scientific literature (Lisboa et al., 2018; Norris et al., 2015).

For instance, classifying solid impurity in raw sugarcane can be performed with chemometric techniques, such as soft independent modeling of class analogy (SIMCA), partial least squares discriminant analysis (PLS-DA), and k-nearest neighbors (kNN) by using the conversion of digital images in color histograms (Guedes and Pereira, 2019). The content of raw sugarcane between 85 and 100 wt% was accurately classified. According to approximately 0.97 of receiver operating characteristic (ROC) area curves for sensibility and specificity using PLS-DA and 1 for SIMCA and kNN. Although these results were promising, the average color values were also tested with no successful results.

In this sense, it is possible to develop a faster computational method using another strategy as the artificial neural network (ANN) model and the averaged color values. The advantage of the averaged color values from images is that ten color-scale represent the average of the color interval with originally 256 intensities/variables, as follows: R (red), G (green), B (blue), their relative colors (r, g, and b), H (hue), S (saturation), V (value) and L (luminosity), which means less running time for computational tests.

The solid impurity in raw sugarcane was successfully estimated using the ANN model for color image data since the data showed no-linear nature. The parameters computed for the ANN model were very promising, the relative errors were 3%, and the data were highly correlated, with the reference values achieving 0.98 for the training set (Guedes et al., 2020).
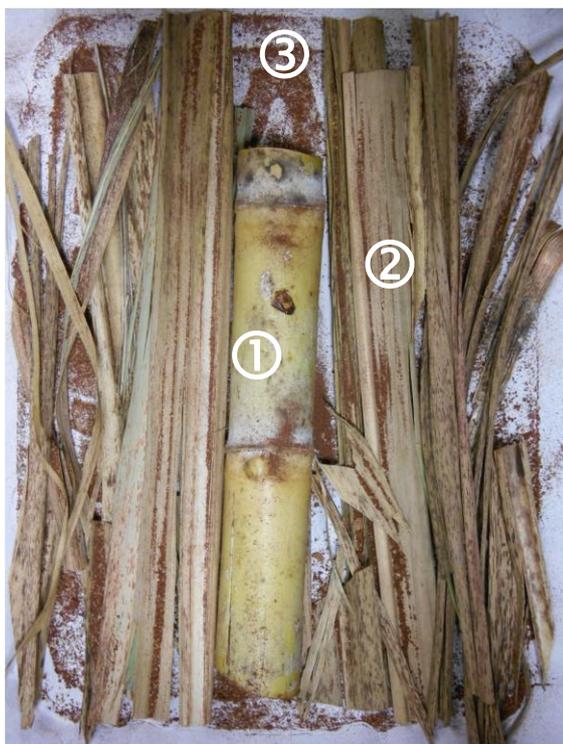
Artificial neural network methods include accurate results, ease of implementation, low computational cost, speed in obtaining results, and the ability to learn through a set of examples and provide consistent responses to new data (Braga et al., 2000; Guedes et al., 2020; Santos et al., 2019).

Therefore, the main goal was to classify raw sugarcane in the presence of solid impurity using the ANN method, as the last part of series of investigations dedicated to this critical issue for sugar mills and biorefineries.

2. Experimental

2.1 Samples and image acquisition

Among sugarcane stalks, vegetal plant parts, and soil, one-hundred forty-six solid mixtures were prepared to acquire digital images, as shown in Fig. 1. Each one was placed in a paper tray (26.5 × 21.5 cm) into a laboratory-made setup (Guedes and Pereira, 2019) composed of a black box, a digital camera Nikon (COOLPIX S3500, Tokyo, Japan) 20.1-megapixel resolution. The images with a 1600 × 1200-pixel size (width × height) and 300 × 300 dpi (dots per inch) resolution were recorded with the tray in a horizontal position. The camera's focal distance was 10 mm, with a maximum aperture of 3.5, and the region of interest (ROI) corresponded to 100% of the original image. During the acquisition of the images, the camera software automatic adjustments were intentionally disabled. Five images were acquired per sample, and the samples were shaken after each image recording to mimic natural conditions at shipments. The same images were converted into colors using an 'imread' function in MATLAB R2020a (MathWorks, Natick, MA, USA). Afterward, the images were converted into color histograms using another function, 'imhist' in MATLAB. The average color values, which were comprised of ten color-scale descriptors: R (red), G (green), B (blue), their relative colors (r, g, and b), H (hue), S (saturation), V (value) and L (luminosity), using a laboratory-made MATLAB code was available in the study of Camargo et al. (2018).

three sets: 70% (102 samples) for training; 15% (22 samples) for validation, to verify that the network is generalizing the information and to interrupt the training before overfitting occurs; the remaining 15% (22 samples) were for the test, independent of the generalization of the neural model.

3. Results and discussion

Figure 2a shows a treemap chart for the color descriptors magnitudes of the digital images. Red, G, and B had interval values: 104-172, 92–159, and 79–159. In the case of L, the counts were from 275 to 483. The relative colors of RGB, represented as r, g, and b, varied from 0.2 to 0.4, H 0.1–0.7, S 0.1–0.4, and V ranged from 0.4 to 0.7 (see the highlighted tiny area by the purple arrow). The representative images for the ideal scenario mean 100% content of sugarcane stalk, as shown in Fig. 2b, and other different situations with more or less solid impurities, far away from biorefineries need, shown in Fig. 2c, d, and e.

Figure 1. Solid mixture sample of sugarcane stalks (85 wt%) denoted as number 1, among vegetal parts (10 wt%) as 2, and soil (5 wt%) as 3.

2.2 Neural model

The development of the neural models was performed using the scaled conjugate gradient backpropagation algorithm (*traincsg*). For this, the MATLAB R2018a software was used, with the 'NNStart' tool available in the software, choosing the pattern recognition app button in which the ANN input layer (with ten neurons representing the ten inputs: R, G, B, H, S, V, r, g, b, and L, number of intermediate layers and an output layer with two neurons were set manually. The number of neurons in the intermediate layer was defined by trial and error to achieve the best classification of raw sugarcane content in solid mixtures.
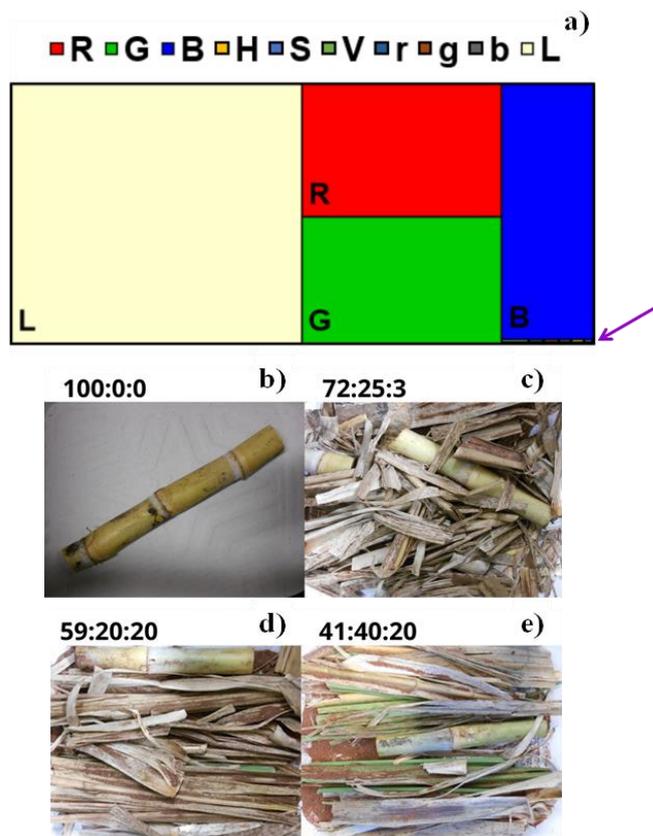
The classification was based on the content of raw sugarcane in wt% denoted as number 1, in the presence of different proportions among green and dry leaves, as number 2, and soil denoted as 3 in Fig. 1. The following division was made: 90–100 wt% designated as class 1 — appropriate (given by binary code 1 0), representing 36 samples; while 41–87 wt% was class 2 — inappropriate (provided by binary code 0 1), representing 110 samples.

The 146 samples were randomly divided using the 'dividerand' function, available in MATLAB, into


Figure 2. (a) Treemap chart of ten color-scale descriptors: R (red), G (green), B (blue), their relative colors (r, g, and b), H (hue), S (saturation), V (value) and L (luminosity) and **(b-e)** Solid mixture samples of sugarcane stalks among vegetal parts, and soil proportions, respectively. Note in the treemap chart

that 'HSV' and 'rgb' are in a tiny area below the right side highlighted by the purple arrow, above B color.

Several architectures were tested to develop the model, varying the number of neurons in the intermediate layer. An artificial neural network with six neurons in the intermediate layer registered the best result with a cross-entropy error of 0.0062 for the validation set. The cross-entropy can measure a classification performance, for which inputs are probability values between 0 and 1. It was observed that ANN could not generalize the problem and classify the samples correctly by increasing the number of neurons in the intermediate layer.

For the training set, the correct classifications of 28 samples of class 1 and 74 samples of class 2 were practicable. For the validation set, no misclassifications for all samples were a remarkable result. Finally, for the test set, all samples were classified as members of their classes. Therefore, all 146 samples achieved a 100% accuracy rate, as shown by the confusion matrices for the training, validation, test sets, and an all-confusion matrix in Fig. 3. Table 1 shows the responses obtained and expected by the ANN for the test set, with five samples of class 1 — appropriate (90–100 wt% of raw sugarcane) and 17 samples of

class 2 — inappropriate (41–87 wt% of raw sugarcane).
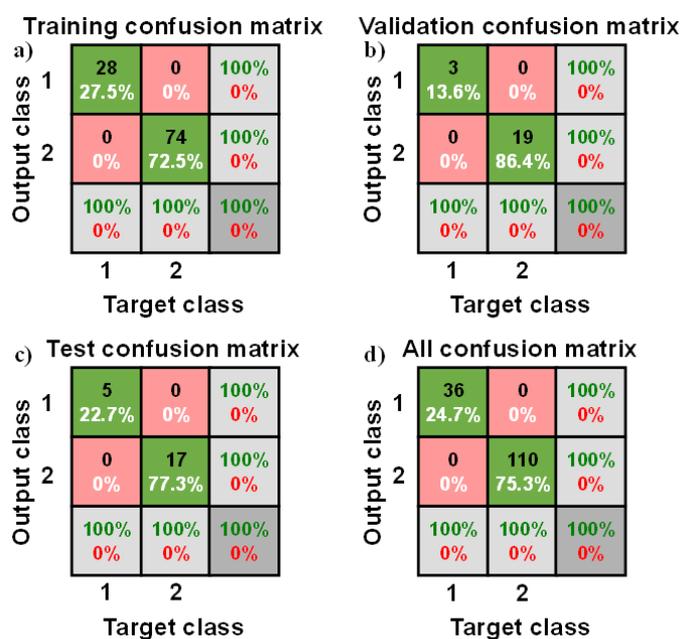
Figure 3. Confusion matrices for sets of training, validation, test, and all results of raw sugarcane content classification in the presence of impurity using an ANN model.

Table 1. Responses were obtained and expected using ANN to classify the raw sugarcane content (wt%) for the test set.

Sample	ANN outputs		Class obtained by the ANN	Targets		Expected class
2	0.0001	0.9999	2	0	1	2
3	0.0006	0.9994	2	0	1	2
16	0.1604	0.8396	2	0	1	2
21	0.0000	1.0000	2	0	1	2
25	0.0000	1.0000	2	0	1	2
28	0.0001	0.9999	2	0	1	2
29	0.0000	1.0000	2	0	1	2
37	0.0000	1.0000	2	0	1	2
51	0.0000	1.0000	2	0	1	2
54	0.0000	1.0000	2	0	1	2
65	0.0001	0.9999	2	0	1	2
68	0.0000	1.0000	2	0	1	2
73	0.0191	0.9809	2	0	1	2
82	0.0001	0.9999	2	0	1	2
86	0.0007	0.9993	2	0	1	2
92	0.0097	0.9903	2	0	1	2
110	0.3197	0.6803	2	0	1	2
111	0.9876	0.0124	1	1	0	1
113	0.9904	0.0096	1	1	0	1
115	0.9993	0.0007	1	1	0	1
133	0.9968	0.0032	1	1	0	1
135	0.9998	0.0002	1	1	0	1

Two other ANN models were also investigated: (i) a model for classifying samples based on the maximum desirable content of vegetal impurity (8 wt%), the range from 0 to 8 wt% designated as class 1 (with binary code 1 0), representing 50 samples, and between 9 and 40 wt% as class 2 (with binary code 0 1), comprising 96 samples; (ii) a model for classification based on the maximum tolerate content of the soil as an impurity (3 wt%), ranging from 0 to 3 was class 1 (with binary code 1 0), with 40 samples and from 4 to 20 wt% as class 2 (with binary code 0 1), a total of 106 samples.

In the ANN model for classifying vegetal impurity content, the best result was 12 neurons in the intermediate layer. The training set was observed among 35 samples of class 1; the neural model misclassified two samples as class 2; among 67 samples of class 2, only one sample was misclassified; that is, the percentage of accurate classifications for the training set was 97.1%. For the validation set, it was observed that the nine samples of class 1 and the 13 samples of class 2 were 100% accurately classified. Finally, for the test set, it was observed that among six samples of class 1, only one sample was misclassified as class 2, and among the 16 samples of class 2, there was one misclassification, that is, the percentage of accurate classifications for the test set was 90.9%. Therefore, for all 146 samples, five were misclassified, representing an average rate of 96.6%.

For classifying the soil content as a solid impurity, 18 neurons in the intermediate layer were the best result. For the training and test sets, eight and three misclassifications were verified, respectively. Therefore, of the 146 samples, 11 were misclassified, representing a 92.5% average rate of accurate classifications.

Thus, the remarkable ANN result is for raw sugarcane content, considering the lowest cross-entropy error, was achieved in this model (0.0062). In contrast, the best models for vegetal parts and soil contents resulted in 0.0160 and 0.0482, respectively.

4. Conclusion

The outstanding result using the ANN method and averaged color values from digital images achieved the lowest cross-entropy errors and 100% of accurate classifications for the content of raw sugarcane, considering the presence of two different types of solid impurity — vegetal plant parts of plants and soil.

Additionally, the ANN running takes a few seconds, and the system of a digital image is an easy-to-use system that can be carried out in any location. Thus, the method can be implemented in sugar cane mills as a screening method of raw sugarcane shipments in the presence of solid impurity as vegetal parts of the plant itself and soil.

Authors' contribution

Conceptualization: Pereira, F. M. V.; Filletti, E. R.
Data curation: Pereira, F. M. V.; Santos, L. J.
Formal Analysis: Santos, L. J.
Funding acquisition: Pereira, F. M. V.; Filletti, E. R.
Investigation: Pereira, F. M. V.; Filletti, E. R., Santos, L. J.
Methodology: Pereira, F. M. V.; Filletti, E. R.
Project administration: Pereira, F. M. V.
Resources: Pereira, F. M. V.; Filletti, E. R.
Software: Filletti, E. R.
Supervision: Filletti, E. R.
Validation: Filletti, E. R.
Visualization: Pereira, F. M. V.; Filletti, E. R.; Santos, L. J.
Writing – original draft: Santos, L. J.
Writing – review & editing: Filletti, E. R.; Pereira, F. M. V.

Data availability statement

The data will be available upon request.

Funding

Acknowledgments

Not applicable.

References

Andrade, D. F.; Guedes, W. N.; Pereira, F. M. V. Detection of chemical elements related to impurities leached from raw sugarcane: Use of laser-induced breakdown spectroscopy (LIBS) and chemometrics, *Microchem. J.* **2018**, *137*, 443–448. https://doi.org/10.1016/j.microc.2017.12.005.

Braga, A. P.; Carvalho, A. C. P. L. F.; Ludermir, T. B. *Redes Neurais Artificiais: Teoria e Aplicações*; Livros Técnicos e Científicos, 2000.

Camargo, V. R.; Santos, L. J.; Pereira, F. M. V. A Proof of Concept Study for the Parameters of Corn Grains Using Digital Images and a Multivariate Regression Model. *Food Anal. Meth.* **2018**, *11*, 1852–1856. https://doi.org/10.1007/s12161-017-1028-6.

Capitán-Vallvey, L. F.; López-Ruiz, N.; Martínez-Olmos, A.; Erenas, M. M.; Palma, A. J. Recent developments in computer vision-based analytical chemistry: A tutorial review. *Anal. Chim. Acta* **2007**, 899, 23–56. https://doi.org/10.1016/j.aca.2015.10.009.

Diniz, P. H. G. D. Chemometrics-assisted color histogram-based analytical systems. *J. Chemom.* **2020**, *34* (12), e3242. https://doi.org/10.1002/cem.3242.

Eggleston, G.; Grisham. M.; Antoine, A. Clarification properties of trash and stalk tissues from sugar cane. *J. Agric. Food Chem.* **2010**, *58* (1), 366–373. https://doi.org/10.1021/jf903093q.

Guedes, W. N.; Pereira, F. M. V. Classifying impurity ranges in raw sugarcane using laser-induced breakdown spectroscopy (LIBS) and sum fusion across a tuning parameter window. *Microchem. J.* **2018**, *143*, 331–336. https://doi.org/10.1016/j.microc.2018.08.030.

Guedes, W. N.; Pereira, F. M. V. Raw sugarcane classification in the presence of small solid impurity amounts using a simple and effective digital imaging system. Comput. Electron. Agric. **2019**, *156*, 307–311. https://doi.org/10.1016/j.compag.2018.11.039.

Guedes, W. N.; Santos, L. J.; Filletti, É. R.; Pereira, F. M. V. Sugarcane stalk content prediction in the presence of a solid impurity using an artificial intelligence method focused on sugar manufacturing. *Food Anal. Methods* **2020**, *13*, 140–144. https://doi.org/10.1007/s12161-019-01551-2.

Lisboa, I. P.; Cherubin, M. R.; Lima, R. P.; Cerri, C. C.; Satiro, L. S.; Wienhold, B. J.; Schmer, M. R.; Jin, V. L.; Cerri, C. E. P. Sugarcane straw removal effects on plant growth and stalk yield. *Ind. Crops Prod.* **2018**, *111*, 794–806. https://doi.org/10.1016/j.indcrop.2017.11.049.

Norris, C. P.; Norris, S. C.; Landers, G. P. A new paradigm for enhanced industry profitability: Post-harvest cane cleaning. In Proceedings of the 37th Conference of the Australian Society of Sugar Cane Technologists, April 28-30, 2015, Bundaberg, Queensland, Australia. Australian Society of Sugar Cane Technologists: Mackay, Australia, 2015.

Pereira, F. M. V.; Bueno, M. I. M. S. Image evaluation with chemometric strategies for quality control of paints. *Anal. Chim. Acta* **2007**, *588* (2), 184–191. https://doi.org/10.1016/j.aca.2007.02.009.

Pereira, F. M. V.; Milori, D. M. B. P.; Pereira-Filho, E. R.; Venâncio, A. L.; Russo, M. S. T.; Martins, P. K.; Freitas-Astúa, J. Fluorescence images combined to statistic test for fingerprinting of citrus plants after bacterial infection. *Anal. Methods* **2011**, *3*, 552–556. https://doi.org/10.1039/c0ay00538j.

Romera, J. P. R.; Barsanelli, P. L.; Pereira, F. M. V. Expeditious prediction of fiber content in sugar cane: An analytical possibility with LIBS and chemometrics. *Fuel* **2016**, *166*, 473–476. https://doi.org/10.1016/j.fuel.2015.11.029.

Santos, M. C.; Nascimento, P. A. M.; Guedes, W. N.; Pereira-Filho, E. R.; Filletti, E. R.; Pereira, F. M. V. Chemometrics in analytical chemistry – an overview of applications from 2014 to 2018. *Eclet. Quim. J.* **2019**, *44* (2), 11–25. https://doi.org/10.26850/1678-4618eqj.v44.2.2019.p11-25.